



A structural equation modeling approach to leveraging the power of extant sentiment analysis tools

Viswadeep Lebakula¹ · Ben Porter² · Megan Stubbs-Richardson³ · Taylor Ray³ · Arthur Cosby³ · Cindy Bethel⁴

Received: 13 June 2024 / Accepted: 7 October 2024
© The Author(s) 2024

Abstract

Machine-derived sentiment analysis has become a pervasive and useful tool to address a wide array of issues in natural language processing. Leading technology companies such as Google now provide sentiment analysis tools (SATs) as readily accessible online products. Academic researchers develop and make available SATs to support the research enterprise. One of the major challenges with SATs is the inconsistencies in results among the various SATs. Consequently, the selection of a SAT for a specific purpose may significantly impact the application. This study addresses the foregoing problem by utilizing structural equation modeling to merge the outputs of SATs to develop a combined sentiment metric without the need for a labeled training dataset. This method is applicable to a wide range of text-based problems, is data-driven, and replicable. It was tested using three publicly available datasets and compared against seven different SATs. The results indicate that as a continuous measure, the proposed method outperformed other SATs in the movie reviews and SemEval datasets, and achieved a tie for first place with IBM Watson on the Sentiment 140 dataset. Also, compared to the published major alternatives, the arithmetic mean solution, this approach performed better across these three datasets.

Keywords Sentiment analysis · Sentiment tools · Natural language processing · Text processing · Combining sentiment analysis tools · Structural equation modeling · Amazon comprehend · Text blob · Vader · Google nlp · Ibm watson · Sentiment 140 · Stanford corenlp

Introduction

Sentiment analysis (SA) is a technique that is used to investigate attitudes, emotions, feelings, opinions, and views expressed in documents [1, 2]. The format of documents can vary from audio, text, and video formats. [3–6]. The term “sentiment analysis” is more commonly used but can also be referred to as “opinion mining”

Extended author information available on the last page of the article

[1–3, 7–10]. Gathering the opinion of individuals is not a new idea. The Greeks used voting to gauge public opinion in the 5th Century B.C. [10]. Research studies on sentiment analysis started during the second half of the 20th century, followed by a surge in publications during the first decade of the 21st century [9, 10]. Over the decades, SA tasks have evolved to analyze multiple document types, domains, and languages. Some applications of SA can be seen in movie reviews, product reviews, restaurant reviews, fake news detection, spam detection, public opinion on government policies, stock price prediction, vote estimation in elections, and to study media tone and polarization [7, 9–15].

Machine Learning Algorithms (MLAs) revolutionized the field of SA. MLAs can learn from big data, which can be processed quickly by High-Performance Computing (HPC). Apart from pre- and post-data processing, SA consists of two main steps: (a) detecting sentiment in the document, and (b) sentiment classification [1, 8]. Issues and complexities regarding SA have not been solved completely, despite SA being a heavily researched problem with only two primary steps, practical functional applications, and a concurrent surge in processing capabilities (e.g., MLA, Big data, HPC). The primary reason for these challenges involves the language itself because processing and understanding natural language (e.g., accounting for words having multiple meanings, sarcasm, and humor) is complex. The complexity further increases with multilingualism.

Three approaches to implement SA are: (a) manual, (b) models, and (c) tools. Manually performing SA is labor-intensive, requires multiple raters to establish reliability, and is not timely. Models and tools can overcome these disadvantages. However, because human interpretation is necessary to determine the true sentiment of a document, the manual approach is essential to generate labeled datasets such as SemEval for training and evaluation of the models and tools [16]. Models were developed by researchers to address their specific research needs. These custom-built models require compute resources and human-generated or labeled datasets to train and validate. In contrast, sentiment analysis tools (SATs) are pre-built, user-friendly applications provided by companies such as Google. While SATs may utilize sophisticated models on the backend, the burden of model development and training is not on the user. Additionally, SATs are advantageous in terms of their ease of use, multilingual support, generalizability across datasets, and eliminates the need for labeled datasets for training (because they have likely already been validated). Researchers and individuals have increasingly adopted SA tools, but there is considerable variability among the results generated by different SATs. Variability among tools suggest that the selection of tools can impact the outcome of a study [17–19].

The current research examined structural equation modeling (SEM), a method commonly used in social sciences, to combine the output from seven SATs into a single metric (i.e., a combined sentiment metric [CSM]). SEM capitalizes on the shared variance between the SAT output to infer sentiment. A minimum of three SATs are required with no upper limit to implement this approach. The major advancement of this method is that, assuming there are validated SATs, it can be applied to any dataset and the result is assumed to be a valid metric of sentiment without the need for a manually labeled dataset. Results suggest that this approach is

equally or more accurate when compared with using any one tool alone and is more effective than the arithmetic mean as performed previously [20]. Additionally, SEM can be used to combine these scores without empirical or subject matter expertise about the relative quality of the component SATs in unique document contexts.

Related research

Based on the algorithmic approach followed, previous research on SA is divided into three categories: (a) machine learning (ML), (b) lexicon, and (c) hybrid [8, 21, 22]. ML is subdivided into supervised learning (SL) and unsupervised learning. SA studies more often employ SL approaches than unsupervised learning [23]. SL approaches consist of five steps: (a) construction of training and test datasets with labels, (b) model building, (c) model training, (d) model testing, and (e) model deployment to perform sentiment analysis on an unseen dataset. ML approaches automatically learn features from data, thereby removing the need of manually coding each word (as is required for lexicon-based approaches), and it provides better accuracy than lexicon models [21]. However, training and testing datasets require resource intensive manual labeling, and corrupted training datasets can compromise the model's effectiveness. Some examples of SL approaches are Support Vector Machines, Neural Networks, and Naïve Bayes [23]. The lexicon-based approaches calculate the sentiment score of the document by assigning values for words from the dictionary [21, 23]. Two prominent lexicon approaches are dictionary-based and corpus-based. Unlike SL, this approach does not require training and testing datasets. However, it requires manual intervention in generating lexicons [21]. A drawback of lexicon-based approaches is that one word may determine the sentiment of a document irrespective of context. Figure 1 presents the sentiment scores from different SATs for a tweet. From Fig. 1, we can see that only the lexicon-based SATs (TextBlob and VADER) generated a negative sentiment score for the tweet because the word ASSAULT is assigned with a negative value in the lexicon. Hybrid approaches are the union of ML and lexicon-based approaches [8, 21]. Hybrid approaches are resilient for a change in the topic domain [8]. They can provide better classification accuracy and precision [24].

Negative Sentiment
word in Lexicon

TWEET : ASSAULT – COLD at NW BROADWAY / NW HOYT ST, PORT
[Portland Police #PP20000145610] 13:12 #pdx911

Amazon Comprehend	Google NLP	IBM Watson	Sentiment 140	Stanford CoreNLP	TextBlob	VADER
0	0	0	0	0	-0.6	-0.6739

Fig. 1 Sentiment scores ranging from -1 (highly negative) to 1 (highly positive) for a tweet by seven tools

There are over 6800 spoken languages with English being the dominant language [25]. The research community has focused on developing SA techniques and lexicons for English [26]. Apart from English, SA is available for other languages such as Mandarin [27], Hindi [14], Spanish [28], and Arabic [29] (the four most spoken languages following English). However, multilingual SA suffers from the scarcity of resources such as lexicons and corpora [26]. One way to overcome resource scarcity is by converting non-English documents into English because of the abundant resources available. The next step is to apply existing tools or models of the English language (such as SentiWordNet) to perform SA. Utilizing this approach, researchers have achieved an accuracy of 66% on the German movie reviews dataset [30]. Most SA techniques can only work on one language, and it is currently not possible for a single model to perform SA on all languages.

Datasets are required to perform model training, validation, and evaluations for SA. Available datasets can be divided into three categories based on the dataset's purpose, the procedure followed for labeling, and the domain types (see Fig. 2). Open source datasets (e.g., STS-Gold) are used in comparative studies to determine the optimal model. Custom-built datasets provide high accuracy. However, they are designed to meet the requirements of a particular study as in [31]. Manual datasets (e.g., SemEval) are labeled by humans whereas, labels for automatic datasets (e.g., Sentiment 140) are generated by using natural language processing techniques or pre-trained models [32]. Datasets are available for different domains such as languages [14, 26–29], medical (e.g., MedBlog) [33], news (e.g., news headlines dataset) [21], social media (e.g., Twitter) [32], and customer reviews (e.g., hotel, movie, restaurant reviews.) [34]. Evaluation metrics for SA techniques are true positive, false positive, true negative, false negative, accuracy, precision, recall, and f1-score [19, 24].

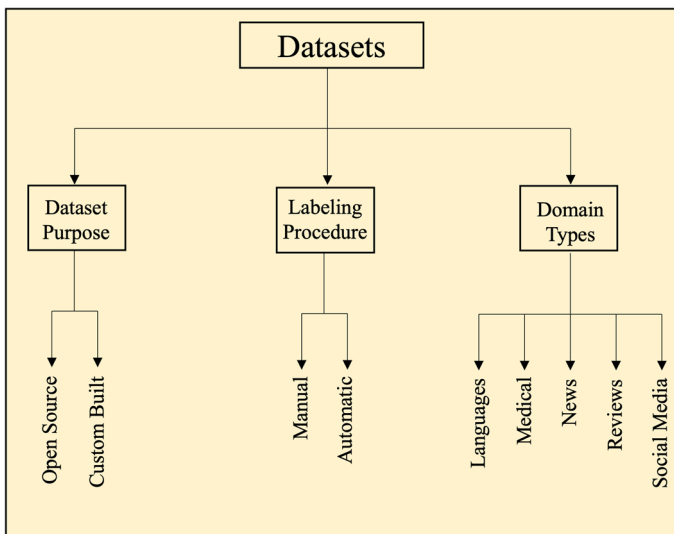


Fig. 2 Datasets classification for sentiment analysis

Different SATs cannot provide an identical sentiment score and this can impact sentiment polarity. Researchers have worked to address this issue by combining the sentiment scores from different SATs. Seven tools (Emoticons, Happiness Index, A Psychometric Scale for Measuring Sentiments on Twitter (PANAS-t), (SailAil Sentiment Analyzer) SASA, SenticNet, SentiStrength, and SentiWordNet) were combined to develop a new tool called the Combined-method [19]. For any given dataset, the Combined-method aims to increase coverage and agreement by analyzing the precision and recall of all of the tools. The authors developed an API iFeel, which enabled researchers to perform comparisons among tools (including the proposed tool). The downside of this method was that it relied on weights constructed from precision and recall, and thus before being applied to a novel data source, would need to have a manually labeled dataset. In a recent study, four SATs (Amazon Comprehend, Google Natural Language, IBM Watson Natural Language Understanding, and Microsoft Text Analytics) scores were combined scores by taking the average [20]. They demonstrated an increased polarity prediction accuracy on a massive open online course (MOOC) dataset. The weakness of this method is that it implicitly assumes that all measures of sentiment are equal. However, prior research has demonstrated that different SATs perform differently with some being better indicators of sentiment.

Sentiment analysis tools

To examine the inconsistencies among SATs, we identified tweets with police keywords from May 1 to 31 of 2020. The keywords to generate this dataset included: police, cops, and sheriff. From these tweets, 300 random tweets were selected each day to create a police tweet dataset with 9,300 police tweets. This dataset allowed for the examination of inconsistencies over time between SATs during a period of changing sentiment toward police in response to George Floyd's death. However, we did not use the police tweets dataset to evaluate the proposed approach (CSM tool) because it requires manual labeling to evaluate the relative performance of various SATs and it is labor intensive. To overcome this drawback, we evaluated the proposed approach with three publicly available datasets.

An overview of the seven SATs used in this research is presented in Table 1. These tools vary in several aspects, such as underlying approach, costs, output, and the number of supported languages. Three (Amazon Comprehend, Sentiment 140, and Stanford CoreNLP) out of seven SATs do not provide output in the range of -1 to 1 . These three SATs were converted into a -1 to 1 scale. For Amazon Comprehend if sentiment = NEUTRAL or MIXED then score = 0, if sentiment = POSITIVE then score = positive likelihood value, and if sentiment = NEGATIVE then score = - negative likelihood value. In the case of Sentiment 140, if output = 4 then score = 1, if output = 0 then score = -1 and if output = 2 then score = 0. For the Stanford CoreNLP if the output = Positive then score = 0.5, if output = Very Positive then score = 1, if output = Negative then score = -0.5 , if output = Very Negative then score = -1 , and if output = neutral then score = 0. After converting the SAT's output into a common scale, they were classified as positive (score > 0),

Table 1 Overview of sentiment analysis tools

Sentiment analysis tools (SATs)	Approach/costs/organization or affiliation	Output/ Multilingual support	Positive tweets	Negative tweets	Neutral tweets
Amazon Comprehend [35]	ML/ Charge varies/ Company	Sentiment and Likelihood for positive, negative, neutral and mixed/ Yes (12 languages)	744	4688	3868
Google NLP [36]	ML/Charge varies/Company	Sentiment score between -1 and 1/ yes (11 languages)	1235	6868	1197
IBM Watson [37]	ML/ Charge varies/ Company	Sentiment score between -1 and 1/ yes (22 languages)	1373	6596	1331
Sentiment 140 [38]	ML/ Free/ University	Sentiment score 4 (positive), 0 (negative), 2 (neutral)/ No	872	2137	6291
Stanford CoreNLP [39]	ML/ Free/ University	Sentiment classes: positive, negative, neutral, very positive and very negative/ Yes (6 languages)	928	2595	5777
TextBlob [40]	Lexicon/ Free / Independent	Sentiment score between -1 and 1/ Yes (2 languages)	3384	3195	2721
VADER [41]	Lexicon + rule-based/ Free/ University	Sentiment score between -1 and 1/ Yes (converts non-English text to English)	3109	4604	1587

Note: Columns 4, 5, and 6 represents total number of tweets classified by sentiment analysis tools as positive, negative, and neutral, respectively

negative (score < 0), and neutral (score = 0). Of note, there is likely a pre-processing procedure programmed into the SATs that provide continuous output given that a score of 0 is unlikely to be common on a continuous scale.

The total positive, negative, and neutral sentiment classification of the police tweet dataset by the SATs are provided in column four of Table 1. Lexicon-based SATs (TextBlob and VADER) predict a higher number of positive tweets than ML-based SATs (Amazon Comprehend, Google NLP, IBM Watson, Sentiment 140, Stanford CoreNLP). Amazon Comprehend identified the lowest number of positive tweets and the highest number was identified by TextBlob. Regarding negative tweets, Google NLP classified the most and Sentiment 140 identified the fewest. Alternatively, regarding neutral tweets, Sentiment 140 classified the most, and Google NLP identified the least.

Often SA studies are conducted over a period of time to determine the change in sentiment toward an entity. Figure 3 depicts the total number of daily negative sentiment police tweets by SATs for a month. From Fig. 3, we can see that the selected SAT can impact the output of time-series studies. If tools are in perfect agreement with each other, then all seven lines in Fig. 3 should overlap, which is not the case. For negative police tweets, VADER and Amazon Comprehend have high agreement with the lowest daily average difference of 2.71. However, this agreement does not hold for a positive and neutral sentiment. Google NLP and Sentiment 140 have the poorest agreement with the highest daily average difference of 152.61.

While all of the SATs demonstrated increasingly negative sentiment toward police following George Floyd’s death, the magnitude of increase varied between

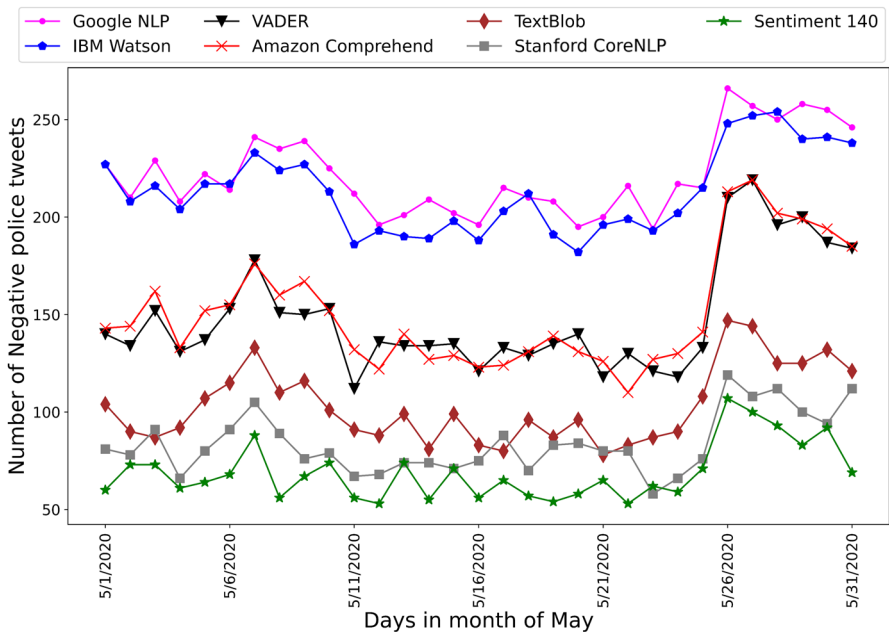


Fig. 3 Daily distribution of negative police tweets by tools for May 2020

the tools with VADER and Amazon Comprehend demonstrating the largest increase. If a researcher used, for example, Sentiment 140 to research this phenomenon, they may determine that negative sentiment did increase, but not to the level of plurality and that sentiment immediately began returning to typical levels. Alternatively, if a researcher used VADER, they would determine that sentiment dramatically increased such that negative sentiment was modal and while a small reduction in negative sentiment was observed, it did not approach normality by the end of May.

This example highlights the difficulties that policymakers and researchers have when applying SATs. Depending on the choice of SAT, two investigations may yield very different conclusions while maintaining the face validity of an increase in negative tweets following George Floyd's death. Further, without manually coding tweets, it is not clear which SAT has the best performance.

Proposed approach

PySpark version 3.1.1 was used for data type conversion, preprocessing, and collecting the outcomes of the SATs. Data manipulation and standard statistical analyses were conducted using SAS software version 9.4. SEM was implemented using Mplus version 8.4 [41].

Structural equation modeling (SEM) was developed in the early 1970 s as a method for using the covariance/variance matrix structure of observed variables to measure unobservable constructs [42]. SEM has been particularly helpful in psychology because it provides a general measurement model for the number of constructs that can only be inferred by symptoms rather than directly observed. For example, depression cannot be directly observed. However, its presence can be inferred from low affect, anhedonia, sleep disturbances, etc. Another way of conceptualizing SEM latent variables is that they are inferred from common variance of the indicators.

Confirmatory Factor Analysis is the most general form of SEM (and the one used in this paper). In this approach, each observed indicator (y_{ni}) is assumed to be determined from a combination of the unmeasured variable (represented by the Greek letter eta, η_i) with a loading (b_{n1} , which is essentially a regression coefficient), and residual variance (refer to equations (1) and (2) below). The only difference between equations (1) and (2) is that they apply to two separate indicators. Residuals (e) and latent variables (η) are assumed to be normally distributed variables with a variance of σ^2 and γ^2 , respectively, and a mean of zero. Because each of the variables contains an identical η_i in their equation, the optimized value of this variable, the loadings, and the residual variance can be estimated to maximize fit in a given dataset. SEM is a particularly effective measurement strategy because the resulting latent variable η is not impacted by the measurement error of the indicators. Variance caused by measurement error is included only in the residuals because such errors are idiosyncratic to each indicator. Of note, while an intercept (b_{n0}) is a component of each variable, it is a constant and does not impact the variance/covariance matrix and so drops from the estimation of η_i .

A maximum likelihood estimation [43] was used to determine the optimal values of these parameters. Metrics of the goodness of fit are available to determine whether a model needs to be modified to increase fit within a dataset [44]. Specifically, common goodness of fit metrics are: comparative fit index (*CFI*), root mean square error of approximation (*RMSEA*), and standardized root mean squared residual (*SRMR*). *CFI* is a measure of fit between the null model (i.e., a poorly fit model without covariances) and the proposed model. *CFI* can be calculated with degrees of freedom (*df*) and chi-square (χ^2) of null (saturated) and proposed (reduced) models; *CFI* > .95 indicates good fit (refer to equation (5)). *RMSEA* is an absolute measure of fit in which values near or less than .05 indicate good fit (refer to equation (6)). *SRMR* is another metric of absolute fit that compares observed covariances (s_{ij}) with estimated covariances ($\hat{\sigma}_{ij}$), values of *SRMR* less than .08 indicate good fit (refer to equation (7)).

$$y_{1i} = b_{10} + b_{11} * \eta_i + e_{1i} \quad (1)$$

$$y_{2i} = b_{20} + b_{21} * \eta_i + e_{2i} \quad (2)$$

$$e_i \sim N(0, \sigma^2) \quad (3)$$

$$\eta_i \sim N(0, \gamma^2) \quad (4)$$

$$CFI = \frac{(\chi_{null}^2 - df_{null}) - (\chi_{proposed}^2 - df_{proposed})}{(\chi_{null}^2 - df_{null})} \quad (5)$$

$$RMSEA = \frac{\sqrt{(x2 - df)}}{\sqrt{[dfN - 1]}} \quad (6)$$

$$SRMR = \sqrt{2 * \sum_{i=1}^p \sum_{j=1}^i [(s_{ij} - \hat{\sigma}_{ij}) / (s_{ii} s_{jj})] / p(p + 1)} \quad (7)$$

Just as it is unknown what is the true level of depression is within an individual, sentiment for a given document is not known (unless rated by humans). Instead, tools estimate the sentiment of a document with presumably varying degrees of success. As such, the results of each tool are informed by, but not identical to, the true sentiment of a document. The degree to which a tool's results are discrepant from true sentiment is due in part to tool-specific measurement error.

As such, sentiment analysis presents an ideal situation to apply SEM to determine the overall sentiment of the document because there are multiple imperfect indicators (tools) of an unobserved construct (true sentiment). CSM is favored

over previous methods for multiple reasons (points 1 to 3), and it also provides extra conveniences (points 4 to 6) that increase its practicality.

1. The loadings of each tool on latent sentiment are not restricted to be equal as is the case when multiple tools are averaged. This allows for tools that better estimate true sentiment to have a greater influence on the resulting latent variable.
2. SEM removes the need for researchers to decide between one tool or another for estimating sentiment and can prevent errors caused by selecting an inappropriate tool. Tools that work well in one context (e.g., with Twitter tweets) may not work well in another (e.g., with Reddit posts). However, given that the most appropriate tool will be most similar to true sentiment, the weights of that tool in an SEM model will be higher than for inappropriate tools.
3. Using SEM removes measurement errors associated with a given tool. By only using common variance to estimate latent sentiment, the estimate is not impacted by systematic problems with individual tools. Averaging will reduce the impact of these problems but will not remove their influence.
4. The maximum likelihood estimation algorithms are incredibly fast to run relative to computationally intensive machine learning algorithms. The equation for the current research took one second to converge.
5. SEM is a well-established procedure and multiple specialized programs (e.g., Mplus, LISREL, AMOS) and non-specialized packages (e.g., R and python packages) exist to estimate SEM models. The output from one program is within a rounding error of the same model estimated in a different program.
6. Finally, SEM models have been expanded to accommodate a wide variety of data inputs and can handle many forms of non-normal data. As such, these models can handle output from tools that produce different types of output (e.g., continuous, ordinal, etc.).

Results

Description of data sources

The proposed approach was tested with three datasets related to two different domains Twitter and Movies. These datasets are widely used and accepted by the research community to perform model evaluations.

- Semantic Evaluation (SemEval) is a series of workshops started in 1998 with a word sense task as the primary focus [44]. Over the decades SemEval evaluations extended to include multiple tasks (e.g., emotion detection, product reviews, and sentiment analysis in Twitter) and languages (e.g., Arabic, Chinese, and Spanish) [16]. For this research, we utilized the task B training dataset from SemEval-2013. This dataset consists of 9,684 tweets with manually classified polarity as positive, negative or neutral.
- The Movie reviews dataset consists of 50,000 IMDB reviews with an equal number of positive and negative reviews [45]. Unlike SemEval, these reviews are coded

automatically using star ratings, which vary from 1 to 10. If a review has a rating seven or higher then it is labeled as positive and labeled as negative if the rating is four or lower. From this dataset, we selected a random sample of 1,500 positive and 1,500 negative reviews.

- Stanford University developed a dataset with 1.6 million tweets to train Sentiment 140 SAT. Similar to the Movie reviews dataset, the Sentiment 140 dataset was not labeled manually but they differ in their automatic labeling procedure. In the Sentiment 140 dataset, tweets were labeled as positive and negative based on positive and negative emoticons, respectively [46]. From this dataset, a set of 3,000 tweets were selected at random with an equal number of positive and negative tweets.

Evaluation metrics

Continuous fit was measured using Spearman's Rho (ρ) as described in this equation:

$$\rho = 1 - \frac{6 \sum d^2}{n(n^2 - 1)} \quad (8)$$

where, d is difference between ranks of an observation from a set of n observations. This correlation coefficient uses rank rather than covariance to determine correlation. As such, it is a non-parametric measure ideal for assessing correlations with ordinal data (i.e., SemEval human-rated values).

Cut points were derived from the continuous latent variable and arithmetic means using Youden's J criteria [47] (i.e., maximizing the sum of sensitivity and specificity) using a subset of documents as a training dataset with the remaining documents used to evaluate agreement. By utilizing these criteria, measures were compared globally using weighted Cohen's kappa [48], which was calculated as shown in the equations below.

$$\text{Weighted Kappa} = 1 - \frac{\sum_{i=1}^k \sum_{j=1}^k w_{ij} x_{ij}}{\sum_{i=1}^k \sum_{j=1}^k w_{ij} m_{ij}} \quad (9)$$

where, k , w_{ij} , x_{ij} , m_{ij} are total codes, weight matrix elements, observed matrix elements, and expected matrix elements, respectively. Negative and positive agreement were assessed using precision ($Prec$), recall (Rec), and f1-score ($f1$) as follows:

$$Prec = \frac{TP}{TP + FP} \quad (10)$$

$$Rec = \frac{TP}{TP + FN} \quad (11)$$

$$f1 = 2 * \frac{Prec * Rec}{Prec + Rec} \quad (12)$$

where, TP, FP, FN, are true positives, false positives, and false negatives, respectively. Documents in Movie reviews and Sentiment 140 datasets were classified as

either positive or negative (i.e., binary classification rather than continuous), only a single cut point can be derived for the CSM tool. Therefore, comparisons between the CSM tool and the tools with three categories would be biased and thus were not conducted.

SATs and CSM tool comparisons

For the Movie reviews dataset, the original model indicated a misfit ($RMSEA = .08$) that was due to the covariance between TextBlob and VADER. This makes logical sense as these two tools are the only lexical-based tools. Adding this covariance resulted in a model with an excellent fit ($CFI = .99$, $RMSEA = .05$, $SRMR = .01$).

For the Sentiment 140 dataset, the original model indicated a misfit ($CFI = .94$, $RMSEA = .11$). This was due to covariances between the residuals of TextBlob and VADER and the residuals of Sentiment 140 and Vader. Adding these covariances resulted in a model with good fit ($CFI = .99$, $RMSEA = .06$, $SRMR = .02$).

For the SemEval dataset, the original model had some indication of misfit ($RMSEA = .08$) and modification indices suggested a residual covariance between TextBlob and VADER. After adding this residual covariance, the model had an excellent fit ($CFI = .99$, $RMSEA = .04$, $SRMR = .01$). The estimated latent sentiment for each document was exported from this model. Using 3000 labeled tweets, using logistic regression, latent sentiment was used to predict two dummy codes identifying negative (vs. neutral or positive) and positive (vs. neutral or negative) tweets. Following each regression, a classification table was created to identify the number of correct and incorrect classifications for potential cutpoints across different values of latent sentiment. The best cutpoint was identified by maximizing Youden's J [47]. Similarly, cut points were determined on the arithmetic average at $-.061$ and $.202$.

Cut points were not developed for the Movie reviews or Sentiment 140 datasets because these only contain positive and negative documents. While a cut point could be developed to distinguish these two types of documents, comparing it to SATs designed to distinguish between three types of documents (positive, neutral, and negative) would not be informative.

Overall comparisons between the SATs and the combined tools are presented in Table 2. As a continuous measure, the CSM tool (we developed using a SEM approach) performed better than all other tools, including the arithmetic average, for both the Movie reviews dataset and the SemEval dataset. However, for the Sentiment 140 dataset both the CSM tool and IBM Watson were tied as the best tools. For SemEval, in terms of the categorical agreement (Weighted Kappa), Amazon Comprehend performed best, but the CSM tool performed second best. These results also demonstrate the substantial variability in the accuracy of the SATs compared to human-rated sentiment.

In comparison, for SemEval dataset, Amazon Comprehend again emerged as an optimal solution for sentiment regarding precision and f1-scores (see Table 3). IBM Watson performed best regarding negative sentiment and Google NLP performed best regarding positive sentiment. The CSM tool consistently performed as one of

Table 2 Overall Measurement of Agreement

Sentiment Analysis Tools (SATs)	Movie reviews (ρ)	Sentiment 140 (ρ)	SemEval (ρ)	SemEval (Weighted Kappa)
CSM Tool (Proposed method)	.77	.62	.72	.62
Arithmetic Average [20]	.72	.60	.68	.53
Amazon Comprehend [35]	.74	.42	.70	.65
Google NLP [36]	.76	.53	.62	.45
IBM Watson [37]	.74	.62	.57	.44
Sentiment 140 [38]	.47	.59	.43	.36
Stanford CoreNLP [39]	.32	.31	.24	.17
TextBlob [40]	.58	.33	.44	.34
VADER [41]	.49	.42	.52	.41

For Each Dataset, Bold Numbers Indicate the Tools with the Highest Agreement Measure

the better tools, having higher recall than Amazon Comprehend and better precision than IBM Watson and Google NLP. It also outperformed the arithmetic average for all metrics, except for the recall associated with negative sentiment.

Selection of the best SAT

In situations where multiple SATs cannot be used to create combined sentiment metric on the whole dataset due to limited budget, SEM can assist the researchers to select the best SAT. Implementing SEM with the paid SATs can become costly, particularly as the number of queries or the size of the dataset grows. Some paid SATs, like Google NLP, either offer free usage or significantly lower fees for the first few queries. So, the researchers can construct an SEM using a sample of the dataset-such as a few hundred randomly selected queries (e.g., tweets or movie reviews)-from a original larger dataset of tens of thousands. The SEM built on this smaller sample can guide the researchers in selecting the best SAT, which can then be applied to the full dataset. This strategy of using SEM on a smaller subset helps reduce expenses for the study.

To create a combined metric, SEM calculates a series of loadings indicating the association between the latent construct and the sentiment tools. These loadings were standardized to make them directly comparable to each other. Loadings of SATs for the three datasets are presented in Table 4. Loadings are highest for Amazon Comprehend in the SemEval dataset, Google NLP and IBM Watson in the Movie reviews dataset, and IBM Watson in the Sentiment 140 dataset, corresponding to the best individual sentiment tool for each dataset as measured in Tables 2 and 3. Outside of identifying the best tool, these loadings generally indicated the relative rankings of each of the remaining SATs. Therefore, researchers would be able to identify a single highly appropriate SAT for a novel population of documents by selecting the tool with the highest loading from the creation of a CSM.

Table 3 Negative and positive sentiment agreement among tools on SemEval dataset

Sentiment analysis tools (SATs)	Precision (negative tweets)	Recall (negative tweets)	f1-scores (negative tweets)	Precision (positive tweets)	Recall (positive tweets)	f1-scores (positive tweets)
CSM tool (Proposed method)	0.58	0.78	0.67	0.77	0.77	0.77
Arithmetic average [20]	0.42	0.84	0.56	0.74	0.75	0.74
Amazon comprehend [35]	0.64	0.74	0.68	0.83	0.74	0.78
Google NLP [36]	0.44	0.74	0.55	0.60	0.84	0.70
IBM Watson [37]	0.35	0.85	0.49	0.67	0.72	0.69
Sentiment 140 [38]	0.52	0.36	0.43	0.66	0.52	0.58
Stanford CoreNLP [39]	0.29	0.35	0.31	0.58	0.25	0.35
TextBlob [40]	0.32	0.44	0.37	0.57	0.72	0.63
VADER [41]	0.47	0.58	0.52	0.58	0.79	0.67

Columns 2, 3, and 4 represents precision, recall, and f1-scores of the Tools in negative sentiment tweets. Columns 5, 6, and 7 represents precision, recall, and f1-scores of the Tools in positive sentiment tweets

Bold Numbers Indicate the Top Three Tools with the Highest Metrics

Table 4 Loadings for different sentiment analysis tools on three datasets

Sentiment analysis tools (SATs)	SemEval	Movie reviews	Sentiment 140
Amazon Comprehend [35]	.84	.84	.53
Google NLP [36]	.81	.91	.83
IBM Watson [37]	.79	.91	.84
Sentiment 140 [38]	.58	.67	.70
Stanford CoreNLP [39]	.34	.39	.49
TextBlob [40]	.56	.76	.61
VADER [41]	.68	.64	.76

For Each Dataset, Bold Numbers Indicate the Tools with the Highest Loadings

CSM tool with and without the best SATs

Given the array of sentiment tools available, it is possible that researchers may not use or have access to what may be the best tools available. To examine the robustness of using the proposed CSM tool, we compared associations with the ground truth from each of the three datasets after excluding the best tool. We also compared this after removing the top three tools leaving only the four lower performing sentiment analysis tools. Results in Table 5 demonstrate small declines in association with the ground truth when excluding the single best tool with the CSM performing better than all but one of the sentiment tools. Larger decreases were observed when excluding the best three tools. We suggest users not to drop multiple best SAT's while calculating CSM as it can reduce the accuracy of CSM. However, in all cases, the CSM performed better than any component tool (i.e., SAT included in estimation). Therefore, CSM can be used to improve even less than ideal sets of SATs.

CSM tool with free SATs

Given the need to perform tasks without the resources to access paid tools, situations may emerge where only free tools are available. Of note, four of the examined sentiment tools (Sentiment 140, Stanford CoreNLP, TextBlob, and VADER) are freely available. We calculated CSM using only these free tools and the results are presented in Table 6. For SemEval and Movie reviews datasets this variable correlated

Table 5 Results of the CSM tool with and without the best sentiment analysis tools

Sentiment analysis tools (SATs)	SemEval (ρ)	Movie reviews (ρ)	Sentiment 140 (ρ)
With all SATs	.72	.77	.62
Without best SAT	.67	.75	.59
Without top three best SATs	.57	.62	.45

For Each Dataset, Bold Numbers Indicate the Tool with the Highest Association with Ground Truth dataset

with the gold standard ($\rho = .57$ for SemEval and $\rho = .62$ for Movie reviews) at a higher level than any of the single free SATs (refer to Tables 2 and 6). In the SemEval dataset, CSM with free SATs ($\rho = .57$) performed equally to one of the paid SATs ($\rho = .57$ for IBM Watson). Similarly, In the Sentiment 140 dataset, this variable correlated with the gold standard ($\rho = .51$) at a higher level than any of the single free sentiment tools except for the Sentiment 140 SAT ($\rho = .59$). Interestingly, in the Sentiment 140 dataset, the correlation between the CSM tool using only free SATs with the gold standard was even higher than one of the paid SATs ($\rho = .42$ for Amazon Comprehend).

Discussion

The current research details the discrepancies between different SATs and how this can lead to different interpretations of research. Also, we classified available SA datasets into three groups based on the dataset's purpose, the procedure followed for labeling, and the domain types. We used SEM, a technique used commonly in the social sciences, to combine multiple measures of sentiment into a unified latent score. Results indicate that this approach was effective in creating a measure that outperformed most measures of sentiment with a fraction of the effort needed to train a unique algorithm and without the necessity of a subject matter expert to select an appropriate pre-made sentiment tool. Additionally, the CSM tool outperformed the arithmetic average. The measure was the best when examining sentiment as a continuous phenomenon, which likely corresponds to the assumed continuous distribution of the latent variable. This approach has several benefits above and beyond other approaches.

1. When using sentiment as a continuous tool, no human-rated dataset is needed. Sentiment can be plotted over time or compared between contexts without any labeled data. Because the CSM tool uses the common variance of each individual tool, the performance of the CSM tool can be assumed to be the best or nearly the best approach. If categorical data (e.g., positive, neutral, or negative sentiment) is needed, a labeled training dataset is needed to derive the specific cut points delineating these categories on the CSM tool. To accomplish this, human raters would need to label a subset of the data according to the desired categories (e.g., positive, neutral, negative) and identify cut points using the methods described in

Table 6 Results of the CSM tool with and without the free sentiment analysis tools

Sentiment Analysis Tools (SATs)	SemEval (ρ)	Movie reviews (ρ)	Sentiment 140 (ρ)
With all SATs	.72	.77	.62
Without free SATs	.57	.62	.51

For Each Dataset, Bold Numbers Indicate the Tools with the Highest Association with Ground Truth dataset

- Sect. 5.3. It may be possible to estimate these without the use of a training dataset using mixture models, but further research is needed to evaluate this method.
2. To improve the performance of SATs, SAT providers such as Amazon, Google, and IBM are constantly updating the SATs. This means that depending on resource allocation by SAT providers the best SAT may change overtime. To get accurate results, in less time, and for less price, researchers and individuals needs to know, what is the best SAT to perform sentiment analysis at a particular time? One approach to answer this question is to evaluate SATs on a dataset with ground truths. This approach has drawbacks; it requires manual labeling to generate ground truths, which is labor intensive, time consuming, and expensive. The CSM approach proposed in this research can overcome these drawbacks by selecting the tool with the highest loadings as a preferred choice.
 3. The use of SEM does not require each of the indicator tools to have comparable weights. In the current research, many indicator tools were poorly associated with the human-rated dataset and as such would ideally not have as much influence on the resulting combined tool. As was observed in the current study, such indicators should have little influence on the latent assessment of sentiment. Furthermore, this determination was made mathematically rather than by researchers, reducing the chance that bias may influence results and increase the replicability. A major added benefit of this is that the method is generalizable to novel contexts. The current research used two disparate types of documents, but any type of document should be optimized by this method.

Some limitations of this study are, firstly, since the output of SATs is the primary input to the CSM tool, the performance of the CSM tool relies on selected SATs. Hence researchers should be cautious at the time of selecting SATs; more inputs of higher quality can result in creating a better CSM tool. Secondly, all the SATs are not free of charge, choosing multiple tools may be expensive, especially for huge datasets. Thirdly, the CSM tool requires at least three SATs to generate the output. The overall time taken (time taken by SATs + time for SEM) to generate a CSM is greater than the time taken by any one of the tools alone. This time can be reduced by processing all the SATs in parallel since the output of one SAT does not depend on another. Finally, the resulting CSM is distributed as a standard normal curve based on the documents used to create it. While higher scores indicate more positive sentiment and lower scores indicate more negative sentiment, zero indicates the average sentiment of the dataset, which may or may not necessarily indicate a neutral sentiment. To create cut points several hundred hand-coded documents are needed.

Conclusion

This research presents a novel method for combining multiple indicators of sentiment together into a single metric (CSM). This procedure shows promise due to its applicability to a wide variety of documents and the removal of the decision point (which sentiment analysis tool to select ?) presented to the researchers. The current

research indicated three uses for the CSM that are either novel or don't require researchers to "guess right" about which SAT to use. First, CSM outperformed other measures when comparing relative sentiment between documents (e.g., does sentiment increase or decrease over time). Second, the CSM performed comparably to the best SATs when categorizing documents based on cut points. Finally, the CSM was able to identify a single appropriate SAT to use on a dataset of unknown attributes.

Acknowledgements This research is not supported under any grant but by the institutional resources of Oak Ridge National Laboratory and Mississippi State University. This material is based upon Cindy Bethel's work supported while serving at the National Science Foundation. Any opinion, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation. This manuscript has been authored by UT-Battelle, LLC, under contract DE-AC05-00OR22725 with the US Department of Energy (DOE). The US government retains and the publisher, by accepting the article for publication, acknowledges that the US government retains a non exclusive, paid-up, irrevocable, worldwide license to publish or reproduce the published form of this manuscript, or allow others to do so, for US government purposes. DOE will provide public access to these results of federally sponsored research in accordance with the DOE Public Access Plan (<https://energy.gov/downloads/doe-public-access-plan>).

Data availability The experiments conducted in the study are based on publicly available data sets that include SemEval (<https://semeval.github.io/>), Movie reviews (<https://www.kaggle.com/datasets/lakshmi25npathi/imdb-dataset-of-50k-movie-reviews>), and Sentiment 140 (<https://www.kaggle.com/datasets/kazanova/sentiment140/data>).

Declarations

Conflict of interest Authors declare no Conflict of interest.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

References

1. Medhat, W., Hassan, A., & Korashy, H. (2014). Sentiment analysis algorithms and applications: A survey. *Ain Shams Engineering Journal*, 5(4), 1093–1113.
2. Asghar, M. Z., Khan, A., Ahmad, S., & Kundi, F. M. (2014). A review of feature extraction in sentiment analysis. *Journal of Basic and Applied Scientific Research*, 4(3), 181–186.
3. Yadav, S. K., Bhushan, M., Gupta, S. (2015). Multimodal sentiment analysis: Sentiment analysis using audiovisual format. In *2015 2nd international conference on computing for sustainable global development (indiacom)*. IEEE, 1415–1419.
4. Kaushik, L., Sangwan, A., Hansen, J. H. (2013). Sentiment extraction from natural audio streams. In *2013 IEEE international conference on acoustics, speech and signal processing*, IEEE, pp. 8485–8489 (2013)
5. Rosas, V. P., Mihalcea, R., & Morency, L.-P. (2013). Multimodal sentiment analysis of Spanish online videos. *IEEE Intelligent Systems*, 28(3), 38–45.

6. Stappen, L., Baird, A., Cambria, E., & Schuller, B. W. (2021). Sentiment analysis and topic recognition in video transcriptions. *IEEE Intelligent Systems*, 36(2), 88–95.
7. Serrano-Guerrero, J., Olivás, J. A., Romero, F. P., & Herrera-Viedma, E. (2015). Sentiment analysis: A review and comparative analysis of web services. *Information Sciences*, 311, 18–38.
8. Alessia, D., Ferri, F., Grifoni, P., Guzzo, T. (2015). Approaches, tools and applications for sentiment analysis implementation. *International Journal of Computer Applications*, 125(3).
9. Ahlgren, O. (2016). Research on sentiment analysis: the first decade. In *2016 IEEE 16th international conference on data mining workshops (ICDMW)*, IEEE, pp. 890–899.
10. Mäntylä, M. V., Gazioti, D., & Kuuttila, M. (2018). The evolution of sentiment analysis—a review of research topics, venues, and top cited papers. *Computer Science Review*, 27, 16–32.
11. Mohan, S., Mullapudi, S., Sammeta, S., Vijayvergia, P., Anastasiu, D. C. (2019). Stock price prediction using news sentiment analysis. In *2019 IEEE fifth international conference on big data computing service and applications (BigDataService)*, IEEE, pp. 205–208.
12. Haselmayer, M., & Jenny, M. (2017). Sentiment analysis of political communication: Combining a dictionary approach with crowdcoding. *Quality & Quantity*, 51, 2623–2646.
13. Hussein, D. M. E.-D. M. (2018). A survey on sentiment analysis challenges. *Journal of King Saud University-Engineering Sciences*, 30(4), 330–338.
14. Sharma, P., Moh, T.-S. (2016). Prediction of Indian election using sentiment analysis on Hindi twitter. In *2016 IEEE international conference on big data (big data)*, IEEE, pp. 1966–1971.
15. Erkantarci, B., Bakal, G. (2023). An empirical study of sentiment analysis utilizing machine learning and deep learning algorithms. *Journal of Computational Social Science*, pp. 1–17.
16. Rosenthal, S., Farra, N., Nakov, P.: SemEval-2017 task 4: Sentiment analysis in Twitter. arXiv preprint [arXiv:1912.00741](https://arxiv.org/abs/1912.00741) (2019)
17. Ahmed Abbasi, A. H., Dhar, M.: Benchmarking twitter sentiment analysis tools. In Proceedings of the ninth international conference on language resources and evaluation (LRECTM14), Reykjavik, Iceland. European Language Resources Association (ELRA) (2014)
18. Jongeling Robbert, D. S., Alexander, S. (2015). Choosing your weapons: On sentiment analysis tools for software engineering research. In *2015 IEEE international conference on software maintenance and evolution (ICSME)*, IEEE, pp. 531–535.
19. Gonçalves, P., Araújo, M., Benevenuto, F., Cha, M. (2013). Comparing and combining sentiment analysis methods. In *Proceedings of the first ACM conference on Online social networks*, pp. 27–38.
20. Pinto, H. L., Rocio, V. (2019). Combining sentiment analysis scores to improve accuracy of polarity classification in MOOC posts. In *Progress in artificial intelligence: 19th EPIA conference on artificial intelligence, EPIA 2019, Vila Real, Portugal, September 3–6, 2019, Proceedings, Part 1 19*, Springer, pp. 35–46.
21. Sadia, A., Khan, F., Bashir, F. (2018). An overview of lexicon-based approach for sentiment analysis. In *2018 3rd International electrical engineering conference (IEEC 2018)*, pp. 1–6.
22. Zhao, X., & Wong, C.-W. (2023). Automated measures of sentiment via transformer-and lexicon-based sentiment analysis (tlsa). *Journal of Computational Social Science*, pp. 1–26.
23. Alshammari, N. F., & AlMansour, A. A. (2019). State-of-the-art review on Twitter Sentiment Analysis. In *2019 2nd International conference on computer applications and information security (ICCAIS)*, IEEE, pp. 1–8.
24. Appel, O., Chiclana, F., Carter, J., Fujita, H. (2018). *Successes and challenges in developing a hybrid approach to sentiment analysis*. Springer.
25. Brown, K., & Ogilvie, S. (2010). *Concise encyclopedia of languages of the world*. Elsevier.
26. Dashtipour, K., Poria, S., Hussain, A., Cambria, E., Hawalah, A. Y., Gelbukh, A., & Zhou, Q. (2016). Multilingual sentiment analysis: State of the art and independent comparison of techniques. *Cognitive Computation*, 8, 757–771.
27. Wang, Z., Joo, V., Tong, C., Chan, D. (2014). Issues of social data analytics with a new method for sentiment analysis of social media data. In *2014 IEEE 6th International conference on cloud computing technology and science*, IEEE, pp. 899–904.
28. Zafra, S. M. J., Valdivia, M. T. M., Camara, E. M., & Lopez, L. A. U. (2017). Studying the scope of negation for Spanish sentiment analysis on Twitter. *IEEE Transactions on Affective Computing*, 10(1), 129–141.
29. Alayba, A. M., Palade, V., England, M., Iqbal, R.: Arabic language sentiment analysis on health services. In *2017 1st international workshop on arabic script analysis and recognition (asar)*, IEEE, pp. 114–118.

30. Denecke, K. (2008). Using sentiwordnet for multilingual sentiment analysis. In *2008 IEEE 24th international conference on data engineering workshop*, IEEE, pp. 507–512.
31. Tiry, E., Oglesby-Neal, A., Kim, K. (2019). *Social media guidebook for law enforcement agencies*. Urban Institute.
32. Saif, H., Fernandez, M., He, Y., Alani, H. (2013). Evaluation datasets for twitter sentiment analysis: A survey and a new dataset, the sts-gold.
33. Denecke, K., & Deng, Y. (2015). Sentiment analysis in medical settings: New opportunities and challenges. *Artificial Intelligence in Medicine*, 64(1), 17–27.
34. Singh, J., Singh, G., & Singh, R. (2016). A review of sentiment analysis techniques for opinionated web text. *CSI Transactions on ICT*, 4, 241–247.
35. Amazon (2021). Amazon Web Services. aws <https://docs.aws.amazon.com/comprehend/latest/dg/comprehend-general.html>.
36. Google (2021). Cloud natural language. Google Cloud <https://cloud.google.com/natural-language>.
37. IBM (2021). Watson Natural Language Understanding. IBM <https://www.ibm.com/cloud/watson-natural-language-understanding>.
38. Alec Go, R. B., Huang, L. (2021). Sentiment140. Sentiment140 <http://help.sentiment140.com/home>.
39. Group, T. S. N. (2021). CoreNLP. CoreNLP <https://stanfordnlp.github.io/CoreNLP/>.
40. Loria, S., et al. (2018). textblob documentation 2(8). <https://buildmedia.readthedocs.org/media/pdf/textblob/latest/textblob.pdf>, Release 0.15
41. Hutto, C., & Gilbert, E. (2014). Vader: A parsimonious rule-based model for sentiment analysis of social media text. In *Proceedings of the international AAAI conference on web and social media*, pp. 216–225.
42. Jöreskog, K. G. (1970). A general method for estimating a linear structural equation system. *ETS Research Bulletin Series*, 1970(2), 41.
43. Myung, I. J. (2003). Tutorial on maximum likelihood estimation. *Journal of Mathematical Psychology*, 47(1), 90–100.
44. Hu, L.-T., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling: A Multidisciplinary Journal*, 6(1), 1–55.
45. Maas, A., Daly, R. E., Pham, P. T., Huang, D., Ng, A. Y., Potts, C. (2011). Learning word vectors for sentiment analysis. In *Proceedings of the 49th annual meeting of the association for computational linguistics: Human language technologies*, pp. 142–150.
46. Go, A., Bhayani, R., Huang, L. (2009). Twitter sentiment classification using distant supervision. CS224N project report, Stanford I(12).
47. Youden, W. J. (1950). Index for rating diagnostic tests. *Cancer*, 3(1), 32–35.
48. Landis, J. R., Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, pp. 159–174.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Authors and Affiliations

Viswadeep Lebakula¹ · Ben Porter² · Megan Stubbs-Richardson³ · Taylor Ray³ · Arthur Cosby³ · Cindy Bethel⁴

✉ Viswadeep Lebakula
lebakulav@ornl.gov
<https://scholar.google.com/citations?user=2ultCaAAAAAJ&hl=en>

Ben Porter

bennett.w.porter@ssrc.msstate.edu

<https://scholar.google.com/citations?user=zwZYVG4AAAAJ&hl=en&oi=ao>

Megan Stubbs-Richardson

megan@ssrc.msstate.edu

<https://scholar.google.com/citations?user=G8jTexcAAAAJ&hl=en>

Taylor Ray

taylor.ray@hardrockdigital.com

Arthur Cosby

arthur.cosby@ssrc.msstate.edu

<https://scholar.google.com/citations?user=NzCFnCgAAAAJ&hl=en>

Cindy Bethel

CBethel@cse.msstate.edu

<https://scholar.google.com/citations?user=Alsv0h8AAAAJ&hl=en>

- ¹ Geospatial Science and Human Security Division, Oak Ridge National Laboratory, 1 Bethel Valley Road, Oak Ridge, TN 37830, USA
- ² Department of Psychology, Mississippi State University, Mississippi State, MS 39762, USA
- ³ Social Science Research Center, Mississippi State University, Mississippi State, MS 39759, USA
- ⁴ Computer Science and Engineering, Mississippi State University, Mississippi State, MS 39762, USA